KEKER VAN NEST & PETERS LATHAM & WATKINS LLP MORRISON FOERSTER

November 22, 2024

Hon. Ona T. Wang
United States Magistrate Judge
Southern District of New York
New York, New York 10007

Re: *Authors Guild et al. v. OpenAI Inc. et al.,*
Case No.: 1:23-cv-08292-SHS-OTW: Resp. to Letter Brief on Models (ECF 270)

This lawsuit, filed by book authors, centers around specific books datasets that were "used to train GPT-3 and 3.5." (*See* ECF 212 at 1 (Plaintiffs' position); *see also* ECF 69 ¶¶ 4, 83-128 (allegations on specific GPT models powering ChatGPT).) OpenAI agreed to provide discovery into these and a number of *other* models—including GPT-1, GPT-2, GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo—and, beyond that, to supplement its response to Interrogatory 11 to identify all models that OpenAI has "made available through its commercial products [*i.e.* ChatGPT and the API] up until the date of the First Consolidated Class Action Complaint." (Ex. A at 5.)

Plaintiffs' letter motion barely mentions these offers. Instead, Plaintiffs demand that OpenAI identify and produce documents on every single model OpenAI ever trained, including hundreds of thousands of research artifacts that OpenAI never used in a product. These models have no conceivable relevance to this lawsuit, and collecting information about them would be a massive forensic exercise that could take years of work. Plaintiffs' request must be denied.

I. OpenAI's Agreed Production Already Includes the Relevant Models

The Court should deny Plaintiffs' request because OpenAI has already provided discovery into the models relevant to Plaintiffs' case—and far more than that. Plaintiffs centered their case on two "internet-based books corpora" broadly referred to as "books1" and "books2," which (Plaintiffs claim) included their published books. (ECF 69 ¶ 112; ECF 212 at 1.) All agree that OpenAI used that data to train the two GPT models generally known as GPT-3 and GPT-3.5 (Ryder Decl. ¶ 6; ECF 212 at 1), which is why OpenAI agreed to collect and produce fulsome discovery on those two models. And OpenAI has already provided a detailed response to Interrogatory 11 that Plaintiffs conceded was "adequate" for the listed models. (Ex. A at 12.)

There is no reason to believe that "books1" or "books2" were used to train any other GPT model, and Plaintiffs have not suggested otherwise. Nonetheless, in an effort to reach a compromise, OpenAI agreed to produce discovery into GPT-1 and GPT-2—models trained on different datasets (e.g., self-published fan fiction and webpages). (Ex. A at 10.) OpenAI also agreed to provide discovery into *subsequent* production models, including GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo, none of which were trained with "books1" or "books2." (*See* Ryder Decl. ¶ 6.) OpenAI also agreed to inspection of the available training data for GPT-1, GPT-2, GPT-3, GPT-3.5, GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo, notwithstanding the fact that Plaintiffs never served a discovery request calling for training datasets. OpenAI offered to further supplement its response

to Interrogatory 11 to identify each and every model made available through its commercial products, *i.e.* ChatGPT and the API, up until the date of the operative complaint. (Ex. A at 5.)

II. Plaintiffs' Request for Hundreds of Thousands of Irrelevant Models Must be Denied

Even though OpenAI's agreements provide everything Plaintiffs need to actually prosecute this case—and then some—Plaintiffs' letter motion barely acknowledges them and instead demands "documents and training data" about new categories of models that Plaintiffs never mentioned, either in their complaint or during the parties' conferrals: **(1)** "research models," **(2)** "GPT-5" and "Orion," which are terms that some have used to refer to models OpenAI may release in the *future*; and **(3)** "models used to power Microsoft's commercial AI products." (ECF 270 at 1.)

Plaintiffs provide no reason to believe that this discovery will uncover evidence that OpenAI has not already agreed to provide. Instead, the sole basis for the demand is a misleading assertion that "the complaint alleges" that OpenAI used "Class Works to train these Additional Models." (*Id.*)[1] As explained below, discovery into these categories of models is wholly unnecessary and massively burdensome. Plaintiffs' latest demand is nothing more than an attempt to impose on OpenAI an untenable and nonsensical discovery obligation that will add nothing to the merits of this case. "Courts routinely deny discovery requests where the discovery sought is not plausibly related to the claims or defenses asserted." *Insured Advocacy Group, LLC v. Spartan Servs. Corp.*, 23-cv-07212 (LJL), 2024 WL 4827251, at *1 (S.D.N.Y. Nov. 19, 2024).

Research Models. OpenAI has been conducting research since it was founded in 2015. (Ryder Decl. ¶ 4.) As an organization that focuses on building AI models, OpenAI has created hundreds of thousands of research artifacts that might each be called a "model." (*Id.*) These research artifacts have been created by hundreds of different OpenAI researchers, many of whom are no longer affiliated with the company. (*Id.*) These research artifacts are just that: research artifacts, not production models like those used to power products like ChatGPT. Producing discovery for these irrelevant research models—which, again, Plaintiffs do not mention in their pleadings and first raised with OpenAI just two days before filing their motion (Ex. A at 4-6)—would serve no purpose. But more to the point: doing so would be an incredibly complicated and burdensome exercise that would tie up OpenAI's resources for months if not years (Ryder Decl. ¶ 4.), interfere with its business operations, and distract from the merits of this case.

GPT-5 and Orion. The Complaint never mentions Orion or GPT-5. Plaintiffs never raised this issue with OpenAI in the discussions leading to this motion. No models bearing the designations

[1] The complaint's generic references to "artificial intelligence models" do not put *all* OpenAI models at issue. To the contrary, Plaintiffs made clear that the "OpenAI[] LLMs" at issue are those referred to as "GPT-N" (ECF 69 ¶ 83) and referenced the GPT-class models hundreds of times throughout the complaint. And the complaint does not include anything at all about "GPT-5," "Orion," or "research models." That alone is reason enough to deny Plaintiffs' request. Fed. R. Civ. P. 26 advisory committee's note to 2000 amendment (noting that a party has "no entitlement to discovery to develop new claims or defenses that are not already identified in the pleadings"); *see also Bruno v. Zimmer, Inc.*, CV15-6129, 2016 WL 4507004, at *5 (E.D.N.Y. Aug. 26, 2016) ("[D]iscovery is not a fishing expedition for Plaintiffs to obtain information to try and create claims that do not already exist.").

"GPT-5" or "Orion" have been completed, much less made available for public use. (Ryder Decl. ¶ 3.) As non-existent and/or non-complete products, they are not appropriate subjects of discovery—particularly absent reason to believe they have anything to do with Plaintiffs' books.

Models that "powered Microsoft's commercial AI products." This issue, which was never raised with OpenAI leading up to this motion, should be denied because OpenAI is not Microsoft, and only *Microsoft* knows what models are used to power *Microsoft's* commercial products. Plaintiffs should seek this discovery from Microsoft. Moreover, the complaint refers only to Microsoft products powered by GPT-3, GPT-3.5, GPT-4, or GPT-4 Turbo (*see* ECF 69 ¶¶ 4, 133) and OpenAI is already providing discovery on those models.

Two other overbroad requests in Plaintiffs' motion warrant particular attention. First, while Plaintiffs purport to be seeking to compel "***training data***" related to these hundreds of thousands of additional models, they do not quote a single request for production of training data. For good reason: Plaintiffs never served a discovery request calling for OpenAI's training datasets. *Vasquez v. Cnty. of Rockland*, 13-civ-5632 (HBP), 2016 WL 413107, at *4 n.4 (S.D.N.Y. Feb. 2, 2016) ("[B]efore a party can ask the court to compel discovery, that party must first serve a discovery request on the adverse party."). In any case, collecting training data for the hundreds of thousands of "models" that Plaintiffs seek to investigate would be incredibly complicated and would take months if not years of work—if it is even possible. (*See* Ryder Decl. ¶¶ 4-5.)

Second, ***Interrogatory No. 10*** is untethered even from the sweeping categories of models described in Plaintiffs' letter brief. This request seeks identifying information for "every version of every LLM, Generative AI system, AI Model, and API product . . . both as it existed (1) after pre-training and before fine-tuning, (2) after fine-tuning, and (3) any fine-tuned versions [of the same]." (ECF 270-4 at 4-5.) This is not only far broader than necessary; it seeks information that would be effectively impossible for OpenAI to gather. OpenAI researchers may create hundreds or thousands of versions of an LLM, including to implement continuous model upgrades, to address minor issues such as bug fixes, and for performance enhancements. (Ryder Decl. ¶ 5.) Only a small handful of these model versions are put into production and referred to as a GPT-class model like GPT-4; the remainder are never made part of a commercial product. (*Id.*) Collecting useful information, including training data and documentation, for each of these model versions would be an extraordinarily complicated and burdensome exercise, and would take months if not years of work, if even possible. (*Id.*) In any case, Plaintiffs have not articulated any justification for identifying every iterative version in the development of any every single model—particularly when OpenAI is already providing extensive discovery into the final version of a model. (*Id.* (versions often reflect "minor issues such as big fixes or performance enhancements").) The request should be denied given the crushing burden such discovery would impose.

Sincerely,

KEKER, VAN NEST & PETERS LLP	LATHAM & WATKINS LLP	MORRISON FOERSTER LLP
/s/ Paven Malhotra	*/s/ Elana Nightingale Dawson*	*/s/ John R. Lanham*
Paven Malhotra	Elana Nightingale Dawson	John R. Lanham